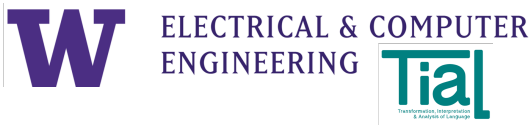
Analysis of Disfluency in Children's Speech

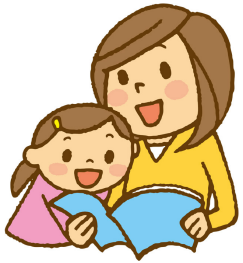
Trang Tran[1], Morgan Tinkler[2], Gary Yeung[2],
Abeer Alwan[2], Mari Ostendorf[1]

[1]*University of Washington*, [2]*University of California Los Angeles*



INTERSPEECH 2020

Why Children's Disfluencies?



- Clinical applications: typical development vs. signs of ASD, ADHD, stuttering
- Non-clinical contexts: understanding language development, signs of uncertainty (and more) in the conversation
- Previous work:
  - disfluency research on adult speech
  - mainly read speech
  - few annotations exist

Contributions

- Novel dataset:

  Disfluency-annotated spontaneous speech from children
- Initial findings from distributional and acoustic analyses
- Automatic disfluency detection results: F1=0.77

Outline

- Background
- Dataset: collection protocol & annotations
- Distributional & acoustic analyses
- Comparison with adult data: distribution & detection
- Summary

Background

- Disfluencies: filled pauses, repetitions, self-corrections

  um so i so i can eat bubblegum ~~every~~ all the time

  - common in spontaneous speech
  - important for spoken language processing
- Related datasets:
  - Child speech: mostly read speech (e.g. Cleuren et al., 2008; Proenca et al., 2015)
  - Adult conversational speech: Switchboard (swbd) & CallHome (callhome)

Dataset Overview

- Part of an effort to develop robots as learning companions
  - Children ages 5–8; 15 female & 11 male
  - 2 interviews, 1 year apart
  - 7 hours of interviews annotated = 1.26 hours of children's speech
  - Teacher prompts child on explanatory discourse tasks:

    "Tell me how you **X**?"; "Why do you **X**?"

    "Now explain to a friend how you **X** and why they should do it"
- Annotate: segment boundaries, fillers, disfluencies (as in swbd), ***plus*** hesitations {H}, partner back channels {PBC}

Annotation Example *(X=brush your teeth)*

Adult (A): Tell me how you brush your teeth.

Child (C): by brushing {H} your tooth {PBC} **//**

A: Okay, anything else you can tell me about how you clean your teeth?

C: {F um} [you + you] get a brush [and then s- + {F um} and then put] it and [some + some] [like + like] just squeeze it **/** and [then + then] you put a little bit of water on it {PBC} **/** and then you brush your teeth **/** and then you spit it out **/**

Segmentation markers: sentence-like unit (SU) (**/**), turn (**//**)
Disfluency mark-up: [reparandum + {interregnum} repair]

Annotator Agreement

- Assessed on 15 interviews (3.7K tokens)
- Boundaries:
  - 4 categories: none, +, /, //
  - {H} mapped to “none” because of low agreement
  - Cohen’s kappa = 0.77
- Disfluencies:
  - 2 categories: 1 (in reparandum) and 0 (not in reparandum)
  - Cohen’s kappa = 0.82

Gender Differences

	overall	female (2x15)	male (2x11)
# tokens	13,568	7,436	6,132
avg. SU len.	6.4	**6.2**	**6.7**
disf. rate	10.1%	**8.5%**	**12.1%**
filler rate	5.0%	**5.4%**	**4.5%**
`uh' rate	0.5%	0.6%	0.5%
`um' rate	2.3%	**2.6%**	**1.9%**
frag. rate	0.8%	**1.2%**	**2.5%**

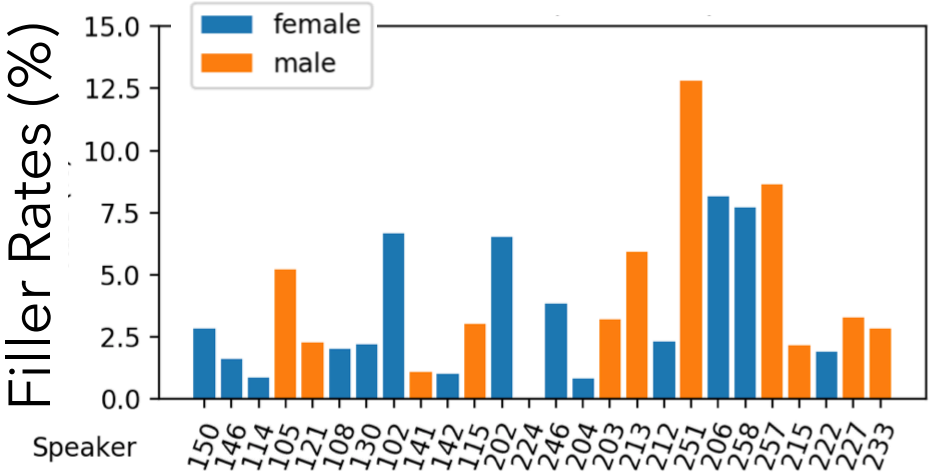
Male vs. female difference is statistically significant at $p < 0.05$

- female children: fewer disfluencies, fragments

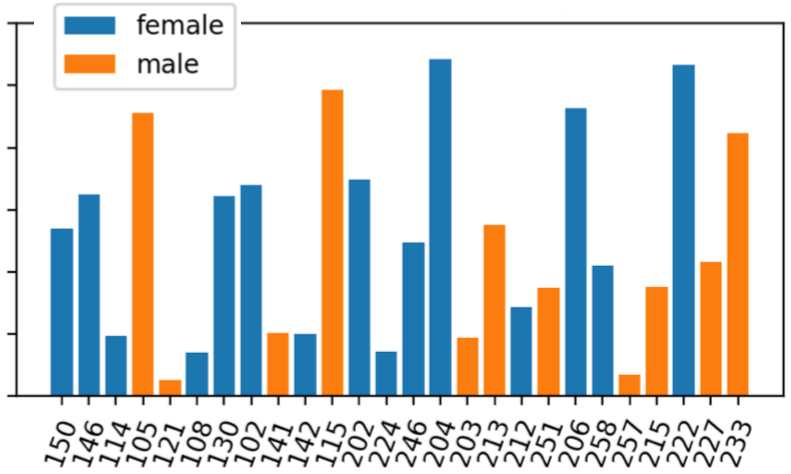
  (similar to adults)
- male children: fewer fillers

  (different from adults)

Session Differences (1 Year Later)



Sess. 1 overall: 3.6%



Sess. 2 overall: 6.0%

- Filler rates: difference is statistically significant ($p < 0.05$)
- Disfluency rates: not significantly different (9.7% & 10.4%)

Task Differences

task	# tokens	avg. SU len	disf. rate	filler rate
teeth 1	2,617	6.3	8.9%	3.2%
colors	2,870	6.3	10.4%	3.9%
animals	1,179	**5.7**	**8.2%**	**7.5%**
teeth 2	3,496	6.6	11.3%	5.7%
blocks	3,406	6.7	10.2%	5.8%

Sess. 1

Sess. 2

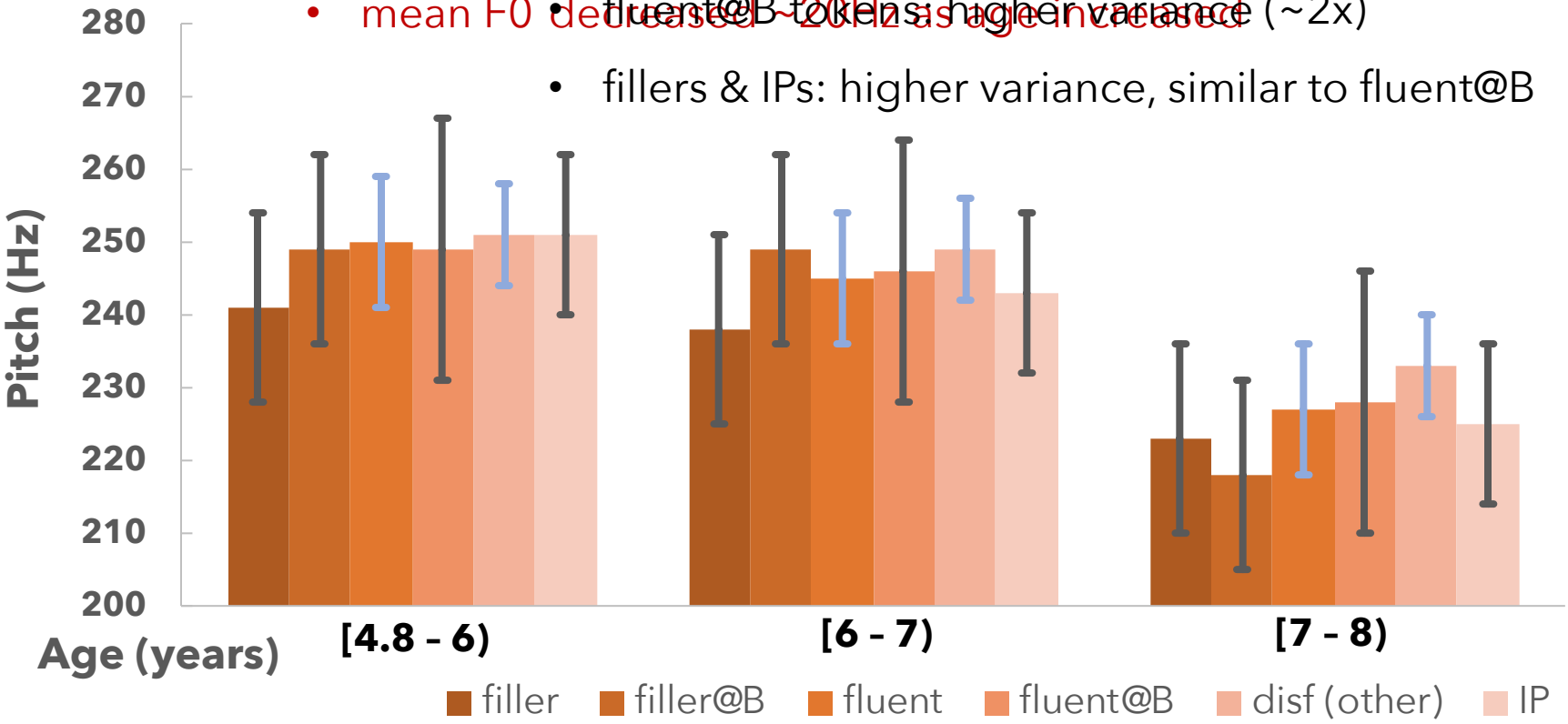
- `animals' task seemed more difficult (children had more questions)
- may have affected other tasks: `teeth 1' vs. `teeth 2'

Acoustic Analysis

- Word-level forced alignment: TDNN trained on TBALL (Kazemzadeh et al., 2015)
  - Largest errors at turn boundaries: excluded from analysis
  - Otherwise: avg. 95 ms error
- Pitch extractor: multi-band summary correlogram (MBSC)-based pitch estimation (Tan and Alwan, 2013)
  - Avg. 15 Hz (6.7% error)
- F0 averaged over word frames, normalized by speaker

Pitch (F0) Findings

- mean F0 decreased ~20Hz as age increased
- fluent@B tokens: higher variance (~2x)
- fillers & IPs: higher variance, similar to fluent@B



Child vs. Adult Speech Corpora

	child	callhome	swbd
# tokens	13,568	43,160	64,944
avg SU len.	**6.4**	7.4	7.5
frag. rate	**1.8%**	1.2%	0.5%
disf. rate	**10.1%**	6.3%	6.2%
filler rate	**5.0%**	3.0%	3.6%
`uh' rate	**0.5%**	0.9%	2.7%
`um' rate	**2.3%**	0.6%	0.5%
avg. ratio repair: reparandum	**0.87**	1.13	1.25

- all differences stat. significant $p < 0.01$
- shorter avg. SUs in children
- higher filler, disf, frag rates in children
- `um' rate > `uh' rate in children
- repair < reparandum in children

Automatic Disfluency Detection

- Disfluency detection system: LSTM-CRF (Zayats & Ostendorf, 2018) trained on SWBD

	child	callhome	swbd
F1 score	0.77	0.66	0.88

- IP detection F1 = 0.73, comparable with previous work on children's speech (Yildrim & Narayanan, 2009)
- Missed disfluencies: longer/more complex

- because [you don't want people to say + when you're talking you don't want people to say] this
- and you can make different colors [at on- + out of + out of] two colors

Summary

- Novel dataset: 1.26 hours of children speech; high-quality disfluency annotations
- Findings on patterns of children's speech:
  - gender differences: disfluency & filler rates
  - disfluency statistics: children exhibit higher disfluency rates and a higher rate for the filled pause "um" (vs. adults)
- Automatic disfluency detection:
  - preliminary result on an adult-speech-trained system (F1 = 0.77)

References

- L. Cleuren, J. Duchateau, P. Ghesquiere, and H. V. hamme, “Children’s oral reading corpus (CHOREC): Description and assessment of annotator agreement,” in LREC, 2008.
- J. Proenca, D. Celorico, S. Candeias, C. Lopes, and F. Perdigao, “Children’s reading aloud performance: a database and automatic detection of disfluencies,” in Proc. Interspeech, 2015.
- S. Yildirim and S. Narayanan, “Automatic detection of disfluency boundaries in spontaneous speech of children using audio-visual information,” IEEE Trans. on Audio, Speech and Language Processing, vol. 17, no. 1, pp. 2 – 12, 2009.
- A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, “TBALL Data Collection: The Making of a Young Children’s Speech Corpus,” in Proc. of EUROSPEECH, 2005, pp. 1581–1584.
- L. N. Tan and A. Alwan, “Multi-Band Summary Correlogram- Based Pitch Detection for Noisy Speech,” Speech Communication, vol. 55, no. 7-8, pp. 841–856, 2013.
- V. Zayats and M. Ostendorf, “Robust cross-domain disfluency detection with pattern match networks,” arXiv preprint arXiv:1811.07236, 2018.

Thank you for watching!

- Dataset:

  `www.seas.ucla.edu/spapl/shareware.html`

- Contact:
  - Trang Tran ttmt001@uw.edu
  - Morgan Tinkler mckeatink@g.ucla.edu
  - Gary Yeung garyyeung@g.ucla.edu
  - Abeer Alwan alwan@ee.ucla.edu
  - Mari Ostendorf ostendor@uw.edu

Backup/Extra Slides

Interview Prompts

- Common questions for tasks:
  - "Tell me how you **X**?"; "Why do you **X**?
  - "Now explain to a friend how you **X** and why they should do it"

- Interview 1:
  - **X** = "brush your teeth" (teeth 1)
  - **X** = "mix paint to make colors" (colors)

- Interview 2:
  - "which animal is the odd one out and why?" (animals)
  - **X** = "count number of cubes" (blocks)
  - **X** = "brush your teeth" (teeth 2)

Annotation Process

- Based on SWBD standard:
  - Turn boundaries: **//** (separation of speaker turns)
  - Sentence-like units: **/** (semantically coherent unit within turns)
  - Filled pauses: {F xx}
  - Disfluencies: [reparandum + {interregnum} repair]
- Extensions:
  - Instructor backchannels: {PBC}
  - Unfilled pauses/duration lengthening: {H}

Analysis Overview

- **Statistical significance in difference** between groups:
  - Length statistics: t-test
  - Rate statistics: Poisson e-test (Krishnamoorthy & Thomson, 2004)
- Group comparisons:
  - female vs. male
  - interview #1 vs. interview #2
  - task X vs. others

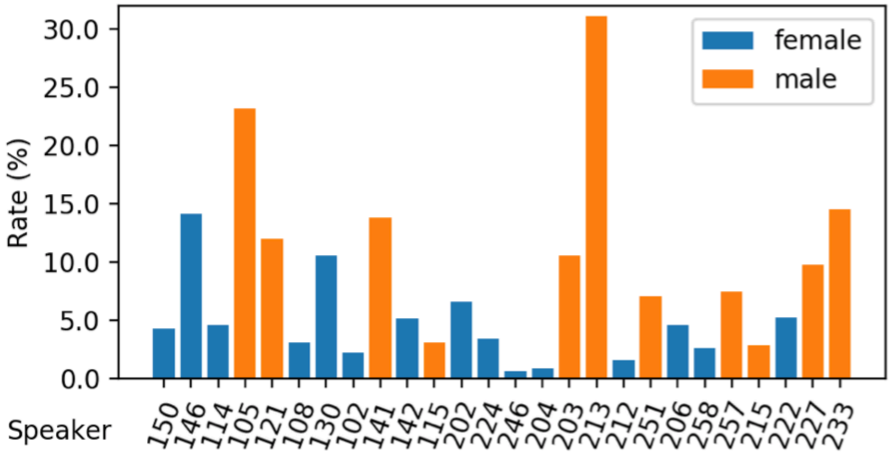
Transcription Statistics: Tasks

task	# tokens	avg. SU len	disf. rate	filler rate
teeth 1	2,617	6.3	**8.9%**	**3.2%**
teeth 2	3,496	6.6	11.3%	**5.7%**
colors	2,870	6.3	10.4%	**3.9%**
animals	1,179	**5.7**	**8.2%**	**7.5%**
blocks	3,406	6.7	10.2%	**5.8%**

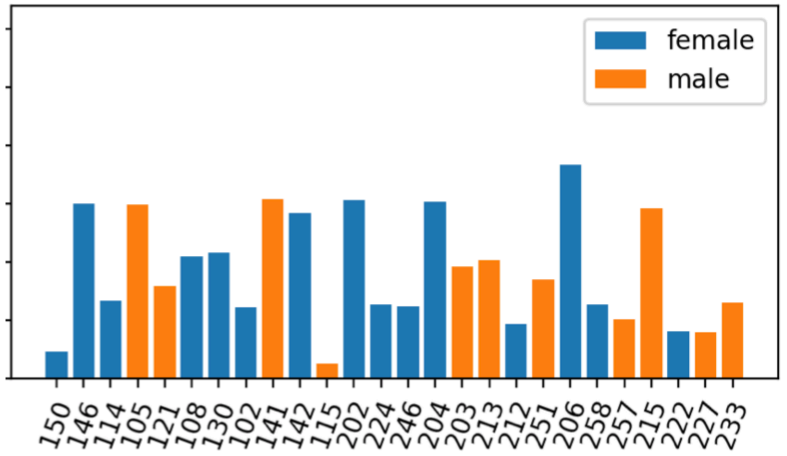
Bold: group difference is statistically significant at $p < 0.05$

- higher disf. and filler rates in second session
- 'animals' task seems most challenging

Disfluency rates between sessions



Sess. 1 overall: 9.7%



Sess. 2 overall: 10.4%

Difference not statistically significant

Token Categories

B = boundaries /, //, +, {PBC}

fillers:

1. not preceding B
2. preceding B

fluent tokens:

3. not preceding B
4. preceding B

disfluent tokens:

5. preceding B (IP)
6. not preceding B

{F um} [it helps me by {F um} + it helps] {H} [kn- + knowing] how many there are /

1 6 6 6 6 2 3 3 5 3 3 3 3 4

Pitch (f0) Findings

Age Category	[4.8-6)	[6-7)	[7-8)
(1) filler	241±15 Hz	238±18 Hz	223±13 Hz
(2) filler @B	249±14 Hz	249±17 Hz	218±13 Hz
(3) fluent	250± 9 Hz	245±10 Hz	227± 9 Hz
(4) fluent @B	249±21 Hz	246±21 Hz	228±18 Hz
(5) IP	251±16 Hz	243±10 Hz	225±11 Hz
(6) other disf.	251±10 Hz	249± 8 Hz	233± 7 Hz

- mean f0 for all categories decreased as age increased
- lower standard deviation for (3) and (6)

- fluent to disfluent region: female f0 increases slightly; male f0 decreases

Pitch (F0) Analysis

- 6 token categories considered, focus on segment boundaries:
  - Fluent tokens with/without boundary
  - Fillers with/without boundary
  - Interruption points (IPs) & other tokens within reparandum
- Findings:
  - mean F0 for all categories decreased (20 Hz avg.) as age increased
  - fluent tokens at SU boundaries have higher variance (2x)
  - fillers and IPs have high variance, similar to fluent tokens at boundary

Automatic Disfluency Detection

- Disfluency detection system: LSTM-CRF (Zayats & Ostendorf, 2018) trained on SWBD

	child	callhome	swbd
F1 score	0.77	0.66	0.88

- IP detection F1 = 0.73, comparable with previous work on children's speech (Yildrim & Narayanan, 2009)

- Missed disfluencies: longer/more complex
  - [[and to + and + and] we have to clean + [if + if you + if] when it's night we have to clean] our teeths
  - because [you don't want people to say + when you're talking you don't want people to say] this
  - and you can make different colors [at on- + out of + out of] two colors

Table 1

Table 1: *Disfluency statistics in the child speech corpus: overall and comparing between genders.* **Bold** *denotes statistically significant difference between genders at* $p < 0.05$.

	overall	female (2x15)	male (2x11)
# tokens	13,568	7436	6132
# turns	2,119	1201	918
avg. SU length	6.4.	**6.2**	**6.7**
disf. rate	10.1%	**8.5%**	**12.1%**
filler rate	5.0%	**5.4%**	**4.5%**
% filler in disf.	12.1%	13.3%	10.2%
'uh' rate	0.5%	0.6%	0.5%
% 'uh' in disf.	16.2%	13.3%	20.7%
'um' rate	2.3%	**2.6%**	**1.9%**
% 'um' in disf.	14.4%	14.4%	14.4%
frag. rate	1.8%	**1.2%**	**2.5%**

Table 2

Table 2: *Disfluency statistics across different tasks.* **Bold** *denotes statistically significant difference between the group and the rest of the groups at* $p < 0.05$.

	teeth 1	teeth 2	colors	animals	blocks
# tokens	2617	3496	2870	1179	3406
# turns	416	532	453	206	512
SU len.	6.3	6.6	6.3	**5.7**	6.7
disf. rate	**8.9%**	11.3	10.4%	**8.2%**	10.2%
filler rate	**3.2%**	**5.7%**	**3.9%**	**7.5%**	**5.8%**
frag. rate	2.0%	1.8%	2.2%	1.2%	1.6%

Table 3

Table 3: *Average mean and standard deviation of f_0 (Hz) for each token category, separated by age.*

Word Category	[4.8-6)	[6-7)	[7-8)
filler	241±15	238±18	223±13
filler@boundary	249±14	249±17	218±13
fluent	250±9	245±10	227±9
boundary	249±21	246±21	228±18
interruption point	251±16	243±10	225±11
within disf.	251±10	249±8	233±7

Table 4

Table 4: *Disfluency statistics across 3 datasets.* **Bold** *denotes statistically significant difference between child speech and adult speech at* $p < 0.01$.

	Child	CallHome	Swbd
# tokens	13,568	43,160	64,944
# turns	2,119	5,869	8,604
avg. SU length	**6.4**	7.4	7.5
disf. rate	**10.1%**	6.3%	6.2%
'uh' rate	**0.5%**	0.9%	2.7%
'um' rate	**2.3%**	0.6%	0.5%
frag. rate	**1.8%**	1.2%	0.5%

Table 5

Table 5: *Average statistics of repair and reparandum lengths in 3 datasets.* **Bold** *denotes statistically significant difference between child speech and adult speech at* $p < 0.01$.

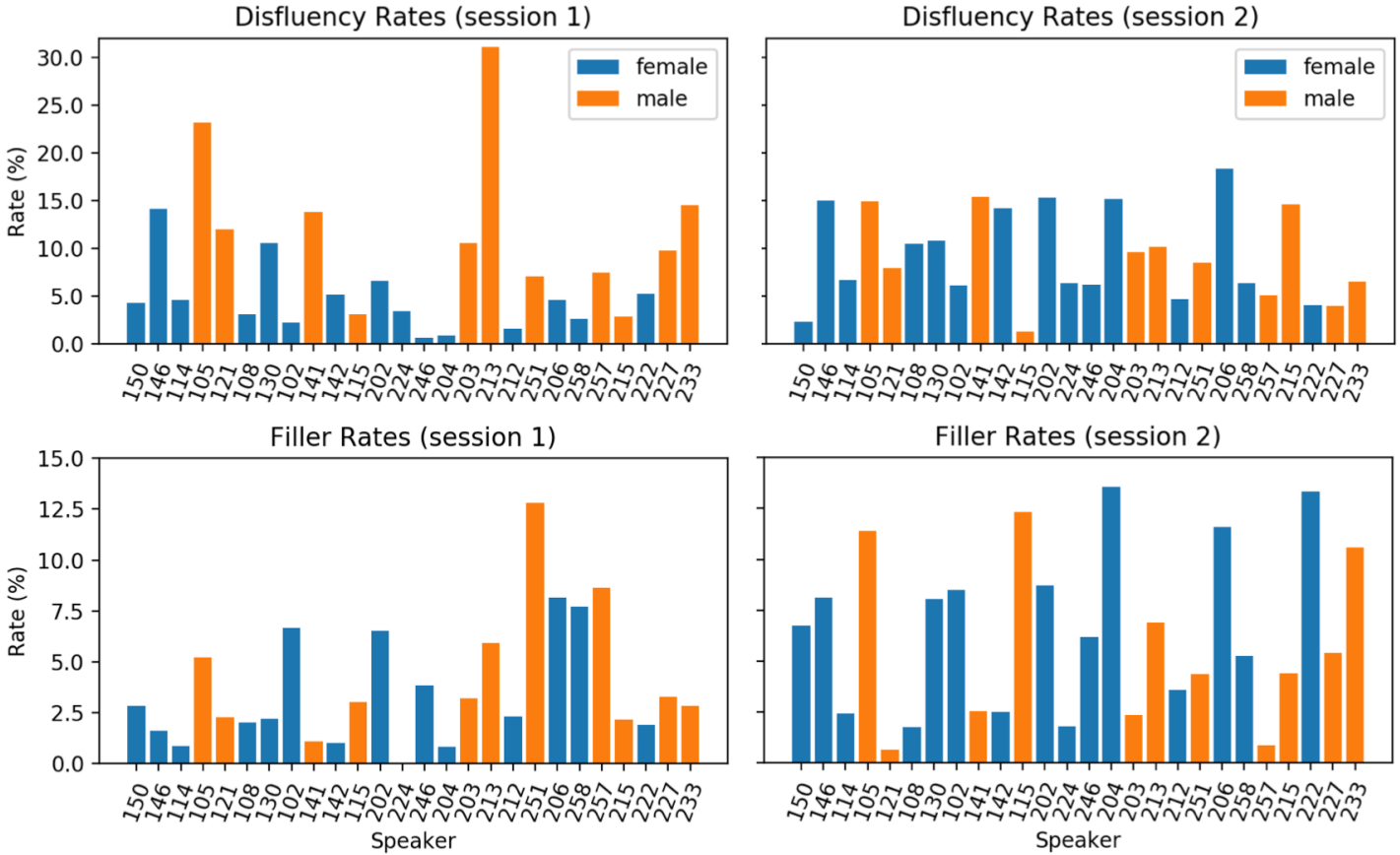
	Child	CallHome	Swbd
# of disfluent regions	525	1068	2159
# non-nested disfluencies	474	922	1923
mean repair length	**1.71**	2.04	1.90
mean reparandum length	**2.46**	2.11	1.59
mean repair:reparandum ratio	**0.87**	1.13	1.25

Table 6

Table 6: *Disfluency detection scores across 3 datasets*

Measure	Child	CallHome	Swbd
precision	0.85	0.66	0.93
recall	0.70	0.66	0.83
F1	0.77	0.66	0.88

Figure 1


Disfluency Rates (session 1)
Disfluency Rates (session 2)
Filler Rates (session 1)
Filler Rates (session 2)
female
male
Rate (%)
Speaker
150 146 114 105 121 108 130 102 141 142 115 202 224 246 204 203 213 212 251 206 258 257 215 222 227 233

Task Differences

- Session 2 `animals' task (*"which is the odd one out?"*) seemed to be more difficult (children asked more questions). It had:
  - higher filler rate (7.5%).
  - lower disfluency rate (8.2%)
  - shorter SUs (5.7 tokens)
- This task may have affected other tasks: differences in the tooth brushing results

`teeth'	disf. rate	filler rate
1	8.9%	3.2%
2	11.3%	5.7%

Contributions

- Novel dataset:

  Disfluency-annotat

  speech from childr
- Initial findings from

  acoustic analyses
- Automatic disfluen

  F1=0.77